

PAREXEL International

AVEO Pharmaceuticals, Inc.

AV-951-15-303

A Phase 3, Randomized, Controlled, Multi-Center, Open-Label Study to Compare
Tivozanib Hydrochloride to Sorafenib in Subjects With Refractory Advanced Renal Cell
Carcinoma

Statistical Analysis Plan

PAREXEL Project Number: 226044

SPONSOR SIGNATURE PAGE

Approved by:



Michael N. Needle, MD
Chief Medical Officer
AVEO Pharmaceuticals, Inc.

1 Oct 2018
Date

PAREXEL SIGNATURE PAGE

This document has been approved and signed electronically on the final page by the following:

Signatory	
Author	Anne Blair Project Role: Biostatistics Lead
Signatory	
Reviewer	Martin Roessner Project Role: Biostatistics Reviewer
	Beate Gerstbrein, PhD Project Role: Medical Writer Reviewer

TABLE OF CONTENTS

1	INTRODUCTION.....	7
2	STUDY OBJECTIVES.....	8
	2.1 Primary Objective.....	8
	2.2 Secondary Objectives.....	8
	2.3 Tertiary Objective.....	8
3	INVESTIGATIONAL PLAN.....	9
	3.1 Overall Study Design and Plan.....	9
	3.2 Efficacy and Safety Variables.....	10
	3.2.1 Efficacy Variables.....	10
	3.2.2 Safety Variables.....	12
4	STATISTICAL METHODS.....	14
	4.1 Data Quality Assurance.....	14
	4.2 General Presentation Considerations.....	14
	4.3 Study Subjects.....	16
	4.3.1 Disposition of Subjects.....	16
	4.3.2 Protocol Deviations.....	16
	4.4 Analysis Populations.....	17
	4.5 Demographic and Other Baseline Characteristics.....	17
	4.6 Treatment Compliance.....	18
	4.7 Efficacy Evaluation.....	19
	4.7.1 Analysis and Data Conventions.....	19
	4.7.1.1 Multi-center Studies.....	19
	4.7.1.2 Adjustments for Covariates.....	19
	4.7.1.3 Handling of Dropouts or Missing Data.....	20
	4.7.1.4 Multiple Comparisons/Multiplicity.....	21
	4.7.1.5 Interim Analyses.....	21
	4.7.1.6 Examination of Subgroups.....	22
	4.7.1.7 Exploratory Analyses.....	22
	4.7.2 Primary Efficacy Variable – PFS.....	22
	4.7.3 Secondary Efficacy Variables.....	24
	4.8 Safety Evaluation.....	24
	4.8.1 Extent of Exposure.....	24
	4.8.2 Adverse Events.....	25
	4.8.3 Prior and Concomitant Medications.....	27
	4.8.4 Clinical Laboratory Evaluation.....	28
	4.8.5 Vital Signs, Physical Findings and Other Observations Related to Safety.....	29
	4.8.6 Safety Monitoring Committee.....	29
	4.9 Determination of Sample Size.....	30
	4.10 Changes in the Conduct of the Study or Planned Analysis.....	31
5	REFERENCES.....	31

LIST OF ABBREVIATIONS

ADI	Actual Dose Intensity
AE	Adverse Event
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
BMI	Body Mass Index
BOR	Best Overall Response
BUN	Blood Urea Nitrogen
CI	Confidence Interval
cm	centimeter(s)
CMH	Cochran-Mantel-Haenszel
CR	Complete Response
CSR	Clinical Study Report
CT	Computerized Tomography
CTC	Common Terminology Criteria
CTCAE	Common Terminology Criteria for Adverse Events
DBP	Diastolic Blood Pressure
DoR	Duration of Response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
EMA	European Medicines Agency
EU	European Union
FDA	Food and Drug Administration
hCG	human chorionic gonadotropin
HR	Hazard Ratio(s)
ICH	International Council for Harmonisation
IDI	Intended Dose Intensity
IMDC	International Metastatic Renal Cell Carcinoma Database Consortium
INR	International Normalized Ratio
IRR	Independent Radiological Review
ITT	Intent-to-treat
IVRS	Interactive Voice Response System
IWRS	Interactive Web Response System
K-M	Kaplan-Meier
kg	kilogram(s)
LDH	Lactate Dehydrogenase
m	meter(s)
MA	Markedly Abnormal
MedDRA™	Medical Dictionary for Regulatory Activities
mg	milligram(s)
mmHg	millimeter(s) of mercury
MRI	Magnetic Resonance Imaging

NA	North America
NCI	National Cancer Institute
NE	Not Evaluable
NMA	Not Markedly Abnormal
NRP	Neuropilin
OR	Objective Response
ORR	Objective Response Rate
OS	Overall Survival
p-y	person-year(s)
PD	Progressive Disease
PE	Physical Examination
PFS	Progression-Free Survival
PP	Per-Protocol
PR	Partial Response
PT	Preferred Term or Prothrombin Time
PTT	Partial Thromboplastin Time
RCC	Renal Cell Carcinoma
RDI	Relative Dose Intensity
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	Serious Adverse Event
SAF	Safety (population)
SAP	Statistical Analysis Plan
SBP	Systolic Blood Pressure
SD	Stable Disease or Standard Deviation
SI	International System (of Units)
SMC	Safety Monitoring Committee
SOC	System Organ Class
T3	Triiodothyronine
T4	Thyroxine
TEAE	Treatment-Emergent Adverse Event
TKI	Tyrosine Kinase Inhibitor
TSH	Thyroid-stimulating Hormone
US	United States
VEGF	Vascular Endothelial Growth Factor
VEGFR	Vascular Endothelial Growth Factor Receptor
WBC	White Blood Cell
WHO	World Health Organization

1 INTRODUCTION

Approximately 208,500 new cases of kidney cancer are diagnosed in the world each year, of which 40,000 new cases are diagnosed in North America and 63,300 new cases in the European Union (EU). Renal cell carcinoma (RCC) accounts for 80%-85% of all malignant kidney tumors. Advanced RCC is highly resistant to chemotherapy, and interleukin-2 and interferon- α have low levels of anti-tumor activity. Recently, drugs that block the VEGF pathway such as sunitinib, sorafenib, bevacizumab, pazopanib, and axitinib have demonstrated significant anti-tumor activity in Phase 3 trials. As a result, these drugs have become the standard of care for the treatment of subjects with advanced RCC.

Tivozanib hydrochloride (tivozanib; previously known as AV-951 and as KRN951) is a novel and potent vascular endothelial growth factor (VEGF) receptor tyrosine kinase inhibitor (VEGFR TKI) that has demonstrated significant anti-tumor effects in pre-clinical experiments. Other agents used for treatment of RCC such as sunitinib and sorafenib also inhibit multiple tyrosine kinases in addition to the VEGF receptor tyrosine kinase, but lead to off-target toxicities such as fatigue, hand-foot syndrome, stomatitis, and neutropenia. The adverse event (AE) profile of tivozanib, however, demonstrates that it is a selective VEGF tyrosine kinase inhibitor, with minimal off-target toxicities. Based on these findings, tivozanib is expected to exhibit a broad anti-tumor spectrum of activities against various tumors by inhibiting angiogenesis and vascular permeability.

TIVO-1 (AV-951-09-301) – a Phase 3, open-label, randomized, controlled, multi-national, multi-center, parallel-arm trial – was conducted to compare tivozanib to sorafenib in patients with advanced RCC. The design of the study was such that patients with radiological evidence of progressive disease (PD) per investigator/local radiology assessment continued treatment until PD was verified by an independent radiologist, after which they were discontinued from study drug treatment and given the option to receive tivozanib under a separate extension protocol (AV-951-09-902). In this study, superiority was shown for the primary endpoint, progression-free survival (PFS), based upon blinded independent radiological assessment (per Response Evaluation Criteria in Solid Tumors [RECIST] Version 1.0 definition). Additional key efficacy endpoints were statistically significantly improved for tivozanib compared with sorafenib, and forest plots assessing the efficacy of tivozanib vs. sorafenib in multiple subgroups showed consistent benefit of tivozanib over sorafenib. There was a slight difference in overall survival (OS), favoring the sorafenib randomized arm over the tivozanib randomized arm; however, an updated analysis of the second line efficacy data from study AV-951-09-902 has revealed that second line efficacy of tivozanib in those patients who crossed over from sorafenib to tivozanib likely drives the discordance between the PFS and OS endpoints.

Study AV-951-15-303 is a Phase 3, randomized, controlled, open-label, multi-center study comparing tivozanib to sorafenib in subjects with refractory advanced RCC. As described above, tivozanib has demonstrated significant activity in the second line of therapy. Sorafenib was chosen as a comparator for this Phase 3 trial because sorafenib is no longer the most commonly used first line VEGFR inhibitor used in advanced RCC.

This Statistical Analysis Plan (SAP) describes the statistical methods to be used during the reporting and analyses of data collected under protocol AV-951-15-303. The planned analyses identified in this SAP may be included in the clinical study report (CSR), any regulatory submissions, or future manuscripts. Post-hoc exploratory analyses not identified in this SAP may also be performed to further examine study data and will be clearly identified as such in the final CSR.

This SAP is based upon the following study documents:

- Study Protocol, Version 3.0 (July 06, 2016) electronic Case Report Form (eCRF), Version 16.0 (March 24, 2016)

The structure and content of this SAP provides sufficient detail to meet the requirements identified by the Food and Drug Administration (FDA), European Medicines Agency (EMA), and International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)[1].

2 STUDY OBJECTIVES

2.1 Primary Objective

The primary objective of this study is as follows:

- To compare the PFS of subjects with refractory advanced RCC randomized to treatment with tivozanib or sorafenib as assessed by blinded independent radiological review (IRR) of computerized tomography (CT) or magnetic resonance imaging (MRI)

2.2 Secondary Objectives

The secondary objectives of this study are as follows:

- To compare the OS of subjects randomized to treatment with tivozanib or sorafenib
- To compare objective response rate (ORR) and duration of response (DoR) of subjects randomized to treatment with tivozanib or sorafenib
- To compare the safety and tolerability of tivozanib and sorafenib

2.3 Tertiary Objective

The tertiary objective of this study is as follows:

- To explore any relationship between tivozanib and sorafenib drug levels and activity
- To explore any relationship between tivozanib and sorafenib drug levels and adverse events (AEs)

3 INVESTIGATIONAL PLAN

3.1 Overall Study Design and Plan

This is a Phase 3, open-label, randomized, controlled, multi-national, multi-center, parallel-arm study comparing tivozanib to sorafenib in subjects with refractory advanced RCC. The study is designed to compare the PFS, OS, ORR, DoR, safety, and tolerability of tivozanib and sorafenib. Approximately 322 subjects with refractory advanced RCC are planned to participate in this study at approximately 160-190 sites. Once informed consent has been obtained and all eligibility criteria have been met (refer to Sections 6.1 and 6.2 of the study protocol for detailed inclusion/exclusion criteria), subjects will be randomized 1:1 using an Interactive Web Response System (IWRS). They will be randomized to treatment with tivozanib or sorafenib, stratified by International Metastatic Renal Cell Carcinoma Database Consortium (IMDC) risk category (favorable; intermediate; poor) and prior therapy (two prior VEGFR TKIs; a prior checkpoint inhibitor [PD-1 or PD1 L inhibitor] plus a prior VEGFR TKI; a prior VEGFR TKI plus any other systemic agent), and must begin treatment within 14 days after randomization. It is anticipated that enrollment to this study will be completed in approximately 15 months.

Tivozanib will be administered orally, at a dose of 1.5 mg/day, beginning on Day 1 of Cycle 1. Subjects will receive tivozanib once daily for 3 weeks followed by 1 week off study drug (1 cycle = 3 weeks on, 1 week off). One cycle will be defined as 4 weeks of treatment. Cycles will be repeated every 4 weeks. Tivozanib should be taken at least 1 hour before or 2 hours after ingesting any food or other medications.

Sorafenib will be administered orally, at an initial dose of 400 mg (two 200 mg tablets) twice daily continuously, beginning on Day 1 of Cycle 1. One cycle will be defined as 4 weeks of treatment. Cycles will be repeated every 4 weeks. Sorafenib should be taken without food at least 1 hour before or 2 hours after ingesting any food or other medications.

Subjects with documented stable disease or an objective response may continue to receive therapy at the same dose and schedule until disease progression or unacceptable toxicities occur, or if other withdrawal criteria are met. Regardless of the duration of the study treatment, all subjects who develop early progressive disease (PD) per RECIST Version 1.1[2] prior to response evaluation will be considered to have progressed on study. Once discontinued, all subjects are to be followed for evidence of delayed effects of study drug for a minimum of 30 days from the last dose of study drug. After the 30-Day Follow-up visit, subjects will be contacted once every 3 months by the sites to collect long-term survival data. Additional contacts may be made to support key analyses, such as the analysis of overall survival at the time of the final PFS analysis, and the final analysis of OS.

Subjects experiencing unacceptable toxicities will be discontinued from further study treatment. Subjects with radiological evidence of progressive disease (per RECIST Version 1.1) per investigator/local radiology assessment should continue treatment until progressive disease is verified by an independent radiologist. An independent safety monitoring committee (SMC) will monitor the safety data from this study on a periodic basis.

Safety assessments are performed on an ongoing basis. The following procedures and evaluations can be collected at unscheduled visits as needed: electrocardiogram (ECG), physical examination, Eastern Cooperative Oncology Group (ECOG) performance status, hematology, serum chemistry, coagulation parameters, thyroid function tests, urinalysis, pregnancy tests, vital signs, tumor assessment, and survival status. Data from unscheduled assessments of the safety endpoints listed above will be included in data listings but not table summaries. For a complete schedule of planned assessments, please refer to Table 1 and/or Section 7 of the study protocol.

3.2 Efficacy and Safety Variables

3.2.1 Efficacy Variables

Tumor response will be assessed using RECIST v1.1 criteria. Subjects will undergo assessments at screening and every 8 weeks following Cycle 1, Day 1 until radiological disease progression is documented and confirmed by independent radiology review. Evaluable subjects are those who receive at least 2 cycles of study drug.

Best overall response (BOR) is determined once all assessments from randomization until disease progression or death are recorded. In general, it is the best response across all assessments; however, confirmation of complete response (CR), partial response (PR), and stable disease (SD) will also be used in BOR determination. To confirm CR or PR changes in tumor measurements must be confirmed by repeat assessments that should be no less than 4 weeks (28 days) after the criteria for response are first met. To confirm SD, it must have occurred at least 8 weeks from Cycle 1, Day 1; otherwise, BOR will depend on subsequent assessments.

All subjects will be followed until death from any cause (unless the subject is lost to follow-up, withdraws consent for the entire study, or the sponsor terminates the study early). After the 30-Day Follow-up visit, subjects will be contacted once every 3 months by the site to collect long-term survival data. For subjects with ongoing objective response or disease stabilization at the time of treatment discontinuation, or for subjects who discontinue for reasons other than radiological evidence of progressive disease, tumor assessments will continue to be collected every 2 months during the first year (beginning from the date of the last scan performed prior to treatment discontinuation), every 3 months during the second year, and every 6 months thereafter until disease progression is confirmed by independent radiology review or the start of another anti-tumor therapy that includes anti-cancer surgery/procedure, anti-cancer radiotherapy, and anti-cancer systemic therapy.

Primary Efficacy Endpoint

PFS is defined as the time from randomization date to the first date of radiological PD or death date (due to any reason), whichever event is earlier. PFS in months will be calculated as: $(\text{first event date} - \text{randomization date} + 1) * (12/365.25)$. Censoring will occur as follows:

- If there are no tumor assessments after randomization, PFS will be censored on the date of randomization.
- If a subject is known to be alive and progression-free, PFS will be censored on the day of the last tumor assessment.
- If a subject is given anti-tumor treatment (anti-cancer surgery/procedure, anti-cancer radiotherapy, and anti-cancer systemic therapy other than the study treatment) prior to PD or death, PFS will be censored on the date of the last tumor assessment prior to the start date of the anti-tumor treatment.
- If treatment follow-up ceases prior to objective PD, PFS will be censored on the date of the last tumor assessment.

Secondary Efficacy Endpoints

OS is defined as the time from the date of randomization to the date of death (due to any reason). OS in months will be calculated as: $(\text{death date} - \text{randomization date} + 1) * (12/365.25)$. Censoring will occur as follows:

- If there are no tumor assessments after randomization, OS will be censored on the date of randomization.
- If there is no confirmation of death, OS will be censored on the date the subject is last known to be alive.

ORR is defined as the proportion of subjects with best overall response (BOR) of confirmed CR or confirmed PR [i.e. objective response (OR)], according to the RECIST 1.1 among subjects with measurable disease at baseline.

DoR is defined as the time from the first date of OR to first date of radiological PD. This will only apply to subjects with confirmed CR or confirmed PR and will be calculated as: $(\text{first event date} - \text{first OR date} + 1) * (12/365.25)$. Censoring will occur as follows:

- If a subject is known to be alive and progression-free, DoR will be censored on the day following the date of the last tumor assessment.
- If a subject dies of a cause known to be non-cancer related, DoR will be censored on the date of the last progression-free tumor assessment.
- If a subject dies of an unknown cause in the absence of documented PD, DoR will be censored on the date of the last tumor assessment.
- If a subject is given anti-tumor treatment (anti-cancer surgery/procedure, anti-cancer radiotherapy, and anti-cancer systemic therapy other than the study

treatment) prior to PD or death, DoR will be censored on the date of the last progression-free tumor assessment prior to the start date of the anti-tumor treatment.

- If treatment follow-up ceases prior to objective PD, DoR will be censored on the date of last progression-free tumor assessment.

Tertiary Efficacy Endpoint

The relationship between tivozanib and sorafenib drug levels and activity will be explored through tumor response assessment using RECIST 1.1 criteria.

The relationship between and tivozanib and sorafenib drug levels and adverse events will be explored through summaries of adverse events related to the study drugs.

3.2.2 Safety Variables

Safety endpoints include study drug exposure, adverse events, laboratory parameters, vital signs, electrocardiogram (ECG), ECOG performance status, concomitant medications, and physical examination.

Study Drug Exposure

Duration of treatment (i.e. exposure) for both tivozanib and sorafenib is defined as the time from the date of the first dose of study treatment to the date of the last dose of study treatment. Duration of treatment is calculated as: (last dose date – first dose date) + 1.

Adverse Event (AE)

Information regarding the occurrence of AEs will be collected throughout participation in this study, starting from the time the subject signs the informed consent form until a period of 30 days after the last dose of study drug. AEs will be coded by system organ class (SOC) and preferred term (PT) using the Medical Dictionary for Regulatory Activities (MedDRA™) Dictionary (Version 18.0 or later). Severity will be graded according to the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) Version 4.03. AEs leading to death or study treatment discontinuation, treatment-related AEs, serious adverse events (SAEs), and treatment-emergent adverse events (TEAEs) will be considered with special attention.

A TEAE is defined as an AE (identified by PT) that begins or that worsens in severity after at least one dose of study treatment has been administered. That is, an AE is considered treatment-emergent if either of the following is met:

- The event begins on or after the first day of study treatment and before 30 days after the last dose of study treatment plus it was never recorded prior to the start of study treatment
- The event was recorded or began prior to the start of study treatment but increased in NCI CTCAE v4.03 grade on or after the first dose of study treatment

Laboratory Parameters

The following laboratory parameters will be collected in accordance with the schedule of assessments as specified in Table 1 of the study protocol:

- Hematology: complete blood count with differential and platelet count
- Serum Chemistry: sodium, potassium, chloride, bicarbonate (optional), blood urea nitrogen (BUN), creatinine, glucose, total bilirubin, direct bilirubin, aspartate aminotransferase (AST), alanine aminotransferase (ALT), lactate dehydrogenase (LDH), alkaline phosphatase, calcium, magnesium, phosphorus, albumin, amylase, lipase, and total protein
- Coagulation: prothrombin time (PT) and/or international normalized ratio (INR), and partial thromboplastin time (PTT)
- Thyroid Function: Free T4 (thyroxine), and thyroid stimulating hormone (TSH) levels. Total T3 (triiodothyronine) is optional
- Urinalysis: protein, glucose, ketones, urobilinogen, occult blood, microscopic sediment evaluation, and 24-hour urine test (required only when protein is at least 4+)
- Pregnancy: serum β -human chorionic gonadotropin (hCG) for premenopausal, non-sterilized females

Vital Signs

Vital sign assessments will be collected in accordance with the schedule of assessments as specified in Table 1 of the study protocol after a subject has been sitting for 5 minutes. They will include measurements of diastolic and systolic blood pressure (mmHg), heart rate (beats per minute), respiration rate (breaths per minute), and body temperature (Celsius).

Electrocardiogram (ECG)

A 12-lead ECG will be performed in accordance with the schedule of assessments as specified in Table 1 of the study protocol after a subject has been supine for approximately 10 minutes. This assessment will include results categorized as normal; abnormal, not clinically significant; abnormal, clinically significant; or unevaluable.

ECOG Performance Status

In accordance with the schedule of assessments as specified in Table 1 of the study protocol, ECOG performance status will be recorded as follows:

- 0 = Fully active, able to carry on all pre disease performance without restriction.
- 1 = Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work.
- 2 = Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours.
- 3 = Capable of only limited self-care, confined to bed or chair more than 50% of waking hours.
- 4 = Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair.
- 5 = Death

Concomitant Medications

Any medication that a subject takes other than tivozanib or sorafenib during the duration of treatment is considered a concomitant medication. Information regarding the usage of concomitant medications will be collected throughout participation in this study, starting from the time the subject signs the informed consent form until a period of 30 days after the last dose of study drug. For each concomitant medication taken during the study, the generic component will be identified using the World Health Organization Drug (WHO Drug) Dictionary.

Physical Examination

A complete physical examination (PE) will be performed at screening by a physician or staff member who is qualified to perform such examinations. The complete PE will entail a thorough review of all body systems, including weight (kg) and height (cm). The remaining PE assessments will be directed PEs, but must include weight (kg). They will be done in a targeted manner, covering body systems related to the subject's disease or relevant findings in the complete PE.

4 STATISTICAL METHODS

4.1 Data Quality Assurance

All tables, figures, and data listings to be included in the report will be independently checked for consistency, integrity and in accordance with standard PAREXEL procedures.

4.2 General Presentation Considerations

'Baseline' is defined as the last observation on or before Cycle 1 Day 1. Baseline assessments should occur within 7 days of Cycle 1 Day 1. 'End of Treatment' is defined as the last available observation up to 7 days after the last dose of study treatment. 'End of Study' is defined as the last available observation after 'End of Treatment'.

Durations are calculated as the stop date minus the start date plus one.

For elapsed time (e.g. time since the initial diagnosis) if the reference date is on or after the event date, then the elapsed time is the reference date minus the event date plus one. If the reference date is before the event date then the elapsed time is the reference date minus the event date.

Unscheduled assessments (laboratory data, ECG, or vital signs associated with non-protocol clinical visits or obtained in the course of investigating or managing adverse events) will be included in listings, but not summaries. If more than one laboratory value is available for a given visit, the first valid observation will be used in summaries and all observations will be presented in listings. If it is not possible to determine which is the first measurement due to missing times then the average of all measurements for that time point will be used as the value for that time point.

Continuous data will be summarized in terms of the mean, standard deviation (SD), median, minimum, maximum, and number of observations, unless otherwise stated. Continuous data that are expected to be skewed will be presented in terms of the maximum, upper quartile, median, lower quartile, minimum, and number of observations. The minimum and maximum will be reported to the same number of decimal places as the raw data recorded in the database. The mean, median, lower quartile, and upper quartile will be reported to one more decimal place than the raw data recorded in the database. The SD will be reported to two more decimal places than the raw data recorded in the database. In general, the maximum number of decimal places reported shall be four for any summary statistic.

Categorical data will be summarized in terms of the number of subjects providing data at the relevant time point (n), frequency counts, and percentages. Any planned collapsing of categories will be detailed in the corresponding sections of this SAP and within the data displays. Percentages will be presented to one decimal place. Percentages will not be presented for zero counts. Percentages will be calculated using n as the denominator.

Changes from baseline in categorical data will be summarized using shift tables where appropriate.

P-values greater than or equal to 0.001, in general, will be presented to three decimal places. P-values less than 0.001 will be presented as "<0.001" and p-values greater than 0.999 will be presented as ">0.999".

Confidence intervals will be presented to one more decimal place than the raw data.

All report outputs will be produced using SAS[®] version 9.2 or a later version in a secure and validated environment. Each report output will be provided to the sponsor in a single Microsoft Word document.

4.3 Study Subjects

4.3.1 Disposition of Subjects

A clear accounting of the disposition of all subjects who enter the study will be provided, from screening to end of study participation. Summaries of subject disposition presented by treatment arm and overall will be provided as follows:

- Number and percentage of subjects randomized
- Number and percentage of subjects randomized by region
- Number and percentage of subjects treated (i.e. received at least one dose of study drug treatment)
- Number of subjects in the analysis populations
- Number and percentage of subjects who withdrew early from the study (including reasons for early withdrawal)
- Number and percentage of subjects who withdrew early from treatment (including reasons for early withdrawal)

By-subject listings of randomization details, visit dates, and withdrawal details (including reason for discontinuation and duration of treatment prior to discontinuation) will also be provided.

4.3.2 Protocol Deviations

Major protocol deviations are defined as those deviations from the protocol likely to have an impact on the perceived efficacy and/or safety of study treatments. The impact of major protocol deviations on the efficacy and/or safety results will be investigated by assessing the robustness of the study results and conclusions to the choice of analysis population (see Section 4.4 for descriptions of analysis populations), both including and excluding data potentially affected by major protocol deviations.

A subject will be classified as having a protocol deviation if one incurs an event belonging to any one of the following categories: 1) violation of the inclusion/exclusion criteria; 2) taking prohibited medications or treatments; 3) treatment misallocation; 4) violation of discontinuation criteria; 5) error in procedure and tests; and 6) missed visits (including procedures and tests).

Major protocol deviations and any action to be taken regarding the exclusion of subjects are defined in the study protocol deviation specifications. The final determination of major protocol deviations and the exclusion of subjects from any of the analysis populations will be made prior to database lock.

A summary of the number and percentage of subjects with a major protocol deviation by treatment arm as well as by type of deviation will be provided. Also, a by-subject listing of major protocol deviations will be provided.

4.4 Analysis Populations

- The intent-to-treat (ITT) population will consist of all randomized subjects; this population will be used for efficacy analysis, including primary endpoints. Subjects will be analyzed as randomized.
- The per-protocol (PP) population is defined as all randomized subjects who received at least two cycles (8 weeks) of protocol treatment, had no major protocol deviations that will confound the effects of treatment, and met all eligibility criteria. The PP population will be used as a sensitivity analysis for the assessment of the primary efficacy endpoint, and subjects will be analyzed as treated.
- The safety (SAF) population will include all subjects who received at least one dose of either study drug. Subjects will be analyzed as treated.

The efficacy summaries and analyses will be based on the ITT population. In the event that a subject is allocated the incorrect study treatment as per the study randomization list, subjects will be summarized and analyzed ‘as randomized’ (i.e. by randomized treatment arm). In the event that a subject is stratified incorrectly, ‘randomized stratum’ will be used rather than ‘actual stratum’.

The safety summaries and analyses will be based on the SAF population. In the event that a subject is allocated the incorrect study treatment as per the study randomization list, subjects will be summarized and analyzed ‘as treated’ (i.e. by allocated treatment arm).

A by-subject listing of analysis population details will be provided. This listing will be presented by treatment arm and will include: center, subject identifier, inclusion/exclusion flag for each population, and reason for exclusion from each population.

4.5 Demographic and Other Baseline Characteristics

The following demographic and baseline characteristics will be summarized for subjects in the ITT population by treatment arm:

- Age at time of informed consent (in years, as a continuous variable)
- Age categorization (<65, >= 65 years)
- Gender (Male, Female)
- Race (White, Black/African American, Asian, American Indian/Alaska Native, Hawaiian Native/Other Pacific Islander, Unknown, Other, Decline to State)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino, Unknown, Decline to State)
- Geographic region [North America (NA), European Union (EU), Rest of World]
- Weight at screening (in kilograms)

- Height at screening (in centimeters)
- BMI at screening (in kg/m²) calculated using weight and height at screening
- Baseline systolic blood pressure (<=140 mmHg, >140 mmHg)
- Baseline diastolic blood pressure (<=90 mmHg, >90 mmHg)
- IMDC risk category (Favorable, Intermediate, Poor)
- Prior therapy (Two Prior VEGFR TKIs, Prior Checkpoint Inhibitor Plus VEGFR TKI, Prior VEGFR TKI Plus Other Systemic Agent)
- Baseline Eastern Cooperative Oncology Group (ECOG) performance status (0, 1)
- Primary diagnosis:
 - Time from initial diagnosis to randomization (in months)
 - Time from most recent relapse or staging to randomization (in months)
 - Pathological diagnosis (Clear Cell, Clear Cell Component, Other)
 - Diagnosis confirmation method (Histology, Cytology, Other)
 - Stage at screening [Stage IV (metastatic), Local Recurrence]
 - Metastatic site location
 - Type of metastases (Multiple Metastases, Solitary Metastases, No Involvement, Unknown)
- Primary disease:
 - Number of metastatic sites per subject
 - Prior radiotherapy (Yes, No)
 - Prior chemotherapy setting (Neo-adjuvant, Adjuvant, Metastatic/Unresectable Therapy, Other, Unknown)
 - Prior nephrectomy (Complete Nephrectomy, Partial Nephrectomy, Other)

Since age is part of the inclusion criteria, age will be calculated as the number of complete years between a subject's birth date and the date of informed consent.

4.6 Treatment Compliance

Compliance with study treatment will be assessed via relative dose intensity (RDI). Relative dose intensity is defined as 100% times the actual dose intensity (ADI) divided by the intended dose intensity (IDI). Actual dose intensity for a subject is defined as the total dose received over all cycles divided by the sum of the lengths of all cycles in days. The intended dose intensity is the total dose assigned over all cycles divided by the sum of the intended lengths of all cycles in days. The assigned dose for a cycle for tivozanib is 1.5 mg/day taken for 3 weeks with 1 week off of study drug. The assigned dose for a cycle for sorafenib is 400 mg twice a day continuously for 4 weeks. For example, for two cycles of treatment, the intended dose intensity (IDI) for sorafenib would be:

$$400 \text{ mg} * 2 * 28 \text{ days} * 2 / 56 \text{ days} = 800 \text{ mg}$$

Summaries of ADI, IDI, and RDI will be presented by treatment arm across all cycles as well as for each cycle.

A by-subject listing of study treatment compliance data will also be provided, containing the date of randomization, the date of last dose of study treatment, and the calculated RDI.

4.7 Efficacy Evaluation

4.7.1 Analysis and Data Conventions

This study is designed to test for superiority. The null hypothesis will be that there is no difference between tivozanib and sorafenib in median PFS. The alternative hypothesis will be that there is a difference. Symbolically, this is expressed as follows:

$$H_0: mPFS_{\text{tivozanib}} \leq mPFS_{\text{sorafenib}}$$
$$H_a: mPFS_{\text{tivozanib}} > mPFS_{\text{sorafenib}}$$

A one-sided, log-rank test stratified for IMDC risk category (favorable vs. intermediate vs. poor) and prior therapy (two prior VEGFR TKIs vs. a prior checkpoint inhibitor plus a prior VEGFR TKI vs. a prior VEGFR TKI plus any other systemic agent) at a significance level of $\alpha = 0.025$ will be used to test this hypothesis.

Estimates of the hazard ratio (HR) for treatment and the corresponding 95% CI will also be presented. They will be estimated using the stratified Cox proportional hazards model with sorafenib as the reference group and stratification factors IMDC risk category and prior therapy.

4.7.1.1 Multi-center Studies

For the purpose of the summaries and analyses mentioned in this SAP, the term ‘Center’ will be used to define each investigator site.

All centers will be pooled together in the efficacy analyses. If the results reveal impact of the centers, then ad hoc analyses will be performed. In that case, the centers may be grouped by geographic regions (see Section 4.5) and evaluation of the consistency of treatment effects across geographic regions may be performed.

4.7.1.2 Adjustments for Covariates

The potential influences of baseline characteristics and stratification factors on the primary PFS endpoint will be evaluated. These factors include the following:

- Age categorization (<65, >= 65 years)
- Race (White, Non-white)
- Gender (Male, Female)
- Baseline Eastern Cooperative Oncology Group (ECOG) performance status (0, 1)
- Time from initial diagnosis to randomization (< 1 year, >=1 year)
- Geographic region [North America (NA), European Union (EU), Rest of World]

- IMDC risk category (Favorable, Intermediate, Poor)
- Prior therapy (Two Prior VEGFR TKIs, Prior Checkpoint Inhibitor Plus VEGFR TKI, Prior VEGFR TKI Plus Other Systemic Agent)
- Number of metastatic sites involved (1, ≥ 2)
- Baseline systolic blood pressure (≤ 140 mmHg, > 140 mmHg) Baseline diastolic blood pressure (≤ 90 mmHg, > 90 mmHg)

To assess the robustness of study conclusions to the choice of covariates, a sensitivity analysis will be performed.

4.7.1.3 Handling of Dropouts or Missing Data

In general, imputed partial dates will not be used to derive study day, duration (e.g., duration of adverse events), or elapsed time variables. In addition, imputed dates are not used for deriving the last contact date in the overall survival analysis dataset.

Imputed dates will not be displayed in listings unless otherwise stated.

The partial date imputation will follow ADaM conventions. The ADaM approach is to populate the numeric date variables with the imputed date and add a flag variable to the dataset that indicates the level of imputation.

The flag variable can contain the values: blank, 'D', 'M', 'Y'.

blank: indicates that no imputation was done

- D='Day': indicates that the day portion of the date is imputed
- M='Month': indicates that the month and day portions of the date are imputed
- Y='Year': indicates that the entire date (year, month, and day) is imputed

Imputing partial AE and prior/concomitant medication start dates:

- a) If the year is unknown, the date will not be imputed and will be assigned a missing value.
- b) If the month is unknown, then:
 1. If the year matches the first dose date, then impute to the month and day of the first dose date.
 2. Otherwise, assign 'January'.
- c) If the day is unknown, then:
 1. If the month and year match the first dose date, then impute to the day of the first dose date.
 2. Otherwise, assign '01'.

Imputing partial AE and prior/concomitant medication stop dates:

- a) If the year is unknown, the date will not be imputed and will be assigned a missing value.
- b) If the month is unknown, then assign 'December'.
- c) If the day is unknown, then impute to the last day of the month.

Imputing partial death dates:

- a) If the month and/or year are unknown, the death date will not be imputed and the subject will be censored for overall survival at the last known date alive.
- b) If the day is missing, then assign '01'.
- c) If imputing the first day of the month results in a negative overall survival time, the subject will be censored for overall survival at the randomization date.

A subject data listing will be provided showing all subjects with missing values for the primary endpoint. For these subjects, the listing will provide all observed data relating to the primary endpoint (i.e. all measurements recorded prior to the missing value, any measurements recorded after the missing value, important baseline characteristics, the recorded reason for study discontinuation, and the date of study discontinuation). The listing will also provide the imputed date(s) used in the primary analysis.

4.7.1.4 Multiple Comparisons/Multiplicity

There will be no adjustment for multiplicity. PFS with one treatment contrast (tivozanib vs. sorafenib) is the only primary analysis defined for this study. The secondary analyses are simply intended to provide supportive evidence relating to the primary analysis, therefore, all efficacy endpoints will be assessed without adjustment for multiple comparisons.

4.7.1.5 Interim Analyses

An interim analysis will be performed after 128 PFS events have been observed. The interim analysis will look into the study's futility with PFS as the efficacy evaluation endpoint. This utilizes a Lan-DeMets spending-type function with an O'Brian-Fleming stopping boundary and estimated two-sided beta (Type II error) of 1.2% spent during this interim look. A planned futility analysis will also be performed.

This interim and futility analyses will be performed by an independent statistician. The independent statistician will evaluate and interpret the results of the interim and futility analyses and will provide corresponding recommendation. The recommended outputs for this reporting effort will be listed in a separate document.

4.7.1.6 Examination of Subgroups

The uniformity of the treatment effect for the primary efficacy variable will be examined for the following subgroups:

1. Age categorization (<65 years, >=65 years)
2. Gender (Male, Female)
3. Race (White, Non-White)
4. Baseline Eastern Cooperative Oncology Group (ECOG) performance status (0, 1)
5. Time from initial diagnosis to randomization (< 1 year, >=1 year)
6. Geographic region [North America (NA), European Union (EU), Rest of World]
7. IMDC risk category (Favorable, Intermediate, Poor)
8. Prior therapy (Two Prior VEGFR TKIs, Prior Checkpoint Inhibitor Plus VEGFR TKI, Prior VEGFR TKI Plus Other Systemic Agent)
9. Number of metastatic sites involved (1, >=2)
10. Baseline systolic blood pressure (<=140 mmHg, >140 mmHg)
11. Baseline diastolic blood pressure (<=90 mmHg, >90 mmHg)

The homogeneity of the treatment effect across subgroups will be investigated using graphical method showing the median treatment effect and 95% confidence interval (CI) within each subgroup. Summaries of the primary efficacy variable by treatment arm and subgroup will be produced.

4.7.1.7 Exploratory Analyses

In order to evaluate if the timing of post-baseline radiological scan influenced the primary outcome, the time from randomization to post-radiological scan will be calculated and an exploratory analyses will be conducted. Mean and standard deviation of radiological scan time will be presented by treatment arms and cycle time. Long-Rank test will be used to test if cumulative percentages (survival curve) are equal (median time to each cycle from randomization) by treatment arms.

4.7.2 Primary Efficacy Variable – PFS

Primary Analysis on PFS

The primary variable for the assessment of efficacy is PFS, as defined in [Section 3.2.1](#) of this SAP.

PFS based on IRR will be summarized for the ITT population using the Kaplan-Meier method. These summaries will include 1- and 2-year survival probabilities along with corresponding 95% CIs. Graphical displays (i.e. K-M plots), where appropriate, will also be produced.

Median event times and corresponding two-sided 95% CIs for the medians will be provided for PFS. A stratified log-rank test (one-sided, $\alpha=0.025$) as well as hazard ratios and corresponding 95% CIs from a stratified Cox proportional hazard model, both adjusting for IMDC risk category and prior therapy, will be used to compare PFS between the two treatment arms.

When determining PFS, the recorded date of progression (defined as first time at which progression can be declared) will be used. For progression due to the presence of a new lesion, the date of progression is the first date that the new lesion was observed. If multiple assessments based on the sum of target lesion measurements are done at different times, the date of progression is the date of the last observation or radiological assessment of target lesions that shows a predefined increase in the sum of the target lesion measurements.

The robustness of the study results to the choice of baseline covariates will be investigated as described in [Section 4.7.1.2](#). In addition, the homogeneity of the treatment effect for important subgroups will be investigated as described in [Section 4.7.1.6](#).

A by-subject listing of the primary efficacy data will be provided.

Sensitivity Analyses on PFS

Sensitivity Analysis #1 will be done using the investigator assessment of response, adding in as events any clinical PDs specified by the investigator. A clinical PD is defined to be treatment failure not meeting the criteria for progressive disease but considered by the investigator to require removal of the subject from the study. This analysis will be conducted using the ITT and PP populations.

Sensitivity Analysis #2 will consider the initiation of new anti-tumor treatment to be an event. In this analysis, any death or PD will also be events, even those that occur after two or more missed tumor assessments. This analysis will be conducted using only the ITT population.

Sensitivity Analysis #3 will be similar to Sensitivity Analysis #2, with the addition of discontinuation of therapy considered an event as well. Again, this analysis will be conducted using only the ITT population.

Sensitivity Analysis #4 will be conducted to assess the impact of missed or not evaluable (NE) tumor assessments. This analysis will use the IRR assessments and backdate any PD events that occur immediately after missing or NE assessments. If the PD occurs immediately after NE assessment (or series of NE assessments), the PD date will be the date of the first NE assessment preceding the PD. If the PD occurs immediately after a missed assessment (or series of missed assessments), the PD date will be the date of the first missed assessment preceding the PD. This analysis will use the ITT population.

All sensitivity analyses will use the same statistical methods as the primary analysis.

Exploratory Analysis on PFS

An exploratory analysis of PFS based on IRR will be summarized for the SAF population using the Kaplan-Meier method. This analysis will examine PFS for all subjects who were randomized and administered at least one dose of the study drug. This analysis will use the same statistical methods as the primary analysis.

4.7.3 Secondary Efficacy Variables

Overall Survival

The first secondary analysis will compare OS between the two treatment arms by using a stratified log-rank test and the stratification factors included in the primary analysis. The hazard ratio for treatment will be estimated using the stratified Cox proportional hazards regression model.

Summaries showing the 1- and 2-year survival probabilities and corresponding 95% CIs will be provided along with graphical displays (similar to PFS).

Objective Response Rate

The number and proportion of subjects achieving confirmed OR (i.e. confirmed CR or confirmed PR) will be summarized and presented by treatment arm and cycle for the ITT population using a Cochran-Mantel-Haenszel (CMH) test ($\alpha=0.05$) with the same stratification factors used for the PFS analysis. The odds ratio and associated 95% CI will be calculated. If the expected number in at least one cell is too small (<5), then Fisher's exact test will be used.

In addition, the BOR for each subject will be summarized (number and percentage of subjects in each category).

Duration of Response

Summaries and graphical displays for DoR will be similar to those of OS. This analysis, however, will only apply to subjects that have achieved confirmed OR.

4.8 Safety Evaluation

All safety summaries and analyses will be based upon the SAF population as defined in [Section 4.4](#) of this SAP.

4.8.1 Extent of Exposure

Tivozanib will be administered once a day for 3 weeks, beginning on Cycle 1 Day 1 of each 4-week cycle. Sorafenib will be administered twice a day continuously, also beginning on Cycle 1 Day 1 of each 4-week cycle.

The following summaries will be produced by treatment arm:

- Number of cycles started
- Duration of treatment in days
- Total days drug was received across all cycles and for each cycle
- Total dose administered in mg across all cycles and for each cycle
- ADI, IDI, and RDI across all cycles and for each cycle
- Average daily dose administered in mg across all cycles and for each cycle
- Number and percentage of subjects on treatment by cycle
- Number and percentage of subjects with dose reduction, including a breakdown of reasons for dose reduction
- Number and percentage of subjects with dose interruption, including a breakdown of reasons for dose interruption

Average daily dose administered is calculated as the total dose administered divided by the total number of days on treatment. ADI, IDI, and RDI calculations were described in [Section 4.6](#).

A by-subject listing of exposure data will also be provided.

4.8.2 Adverse Events

Adverse events will be coded by system organ class (SOC) and preferred term (PT) using the Medical Dictionary for Regulatory Activities (MedDRA™) Version 18.0 or later.

A TEAE is defined as an AE (identified by PT) that begins or that worsens in severity after at least one dose of study treatment has been administered. Treatment-related TEAEs are those with reasonable causality to study drug marked as “yes” on the eCRF. TEAEs with an outcome of death are those with a grade of 5 or an outcome of “fatal.” TEAEs leading to treatment withdrawal are those with a study drug action taken of “drug withdrawn,” while TEAEs leading to study drug dose reduction or interruption are those with a study drug action taken of “dose reduced” or “drug interruption” (respectively).

Summary tables will be presented for TEAEs by treatment arm. An overall presentation of TEAE information will include the following:

- Number and percentage of subjects with at least one TEAE
- Number and percentage of subjects with at least one TEAE of Grade 3 or higher
- Number and percentage of subjects with at least one treatment-related TEAE
- Number and percentage of subjects with at least one treatment-related TEAE of Grade 3 or higher
- Number and percentage of subjects with at least one serious TEAE

- Number and percentage of subjects with at least one treatment-related serious TEAE
- Number and percentage of subjects with at least one TEAE leading to death
- Number and percentage of subjects with at least one treatment-related TEAE leading to death
- Number and percentage of subjects with at least one TEAE leading to treatment withdrawal
- Number and percentage of subjects with at least one treatment-related TEAE leading to treatment withdrawal
- Number and percentage of subjects with at least one TEAE leading to dose reduction
- Number and percentage of subjects with at least one treatment-related TEAE leading to dose reduction
- Number and percentage of subjects with at least one TEAE leading to dose interruption
- Number and percentage of subjects with at least one treatment-related TEAE leading to dose interruption

TEAE information will also be presented by NCI CTCAE severity grade and by relationship to study treatment. If a subject had more than one occurrence of a TEAE, the most severe grade will be used in the summary tables.

The following summaries will be provided:

- Number and percentage of subjects for each TEAE, categorized by SOC and PT
- Number and percentage of subjects for each TEAE of Grade 3 or higher, categorized by SOC and PT
- Number and percentage of subjects for each TEAE, categorized by SOC, PT, and maximum CTCAE grade
- Number and percentage of subjects for each TEAE, categorized by SOC, PT, and relationship to study drug
- Number and percentage of subjects for each treatment-related TEAE, categorized by SOC and PT
- Number and percentage of subjects for each treatment-related TEAE of Grade 3 or higher, categorized by SOC and PT
- Number and percentage of subjects for each treatment-related TEAE, categorized by SOC, PT, and maximum CTCAE grade
- Number and percentage of subjects for each treatment-related TEAE leading to treatment withdrawal, categorized by SOC and PT

Counts will be by subject, not by event, and subjects are only counted once within each SOC or PT. For tables categorized by maximum CTCAE grade, subjects with multiple events within a particular SOC or PT will be counted under the category of their most severe event within that SOC or PT.

Exposure-adjusted TEAE incidence rates per 100 patient-years (p-y) may also be calculated by treatment arm for preferred terms of interest, as identified by AVEO. If requested, the exposure-adjusted incidence rate (EAIR) per 100 p-y is the number of subjects who experience a TEAE divided by the total subject exposure time multiplied by 100. For subjects who experience a TEAE, the exposure time is from the date of first dose of study treatment to the time of the first occurrence of that TEAE. For subjects who do not experience a TEAE, the exposure time is from the date of first dose of study treatment to the date of last dose of study treatment + 30 days or death date, whichever comes first.

A by-subject listing of all AEs (including non-treatment-emergent events) will be provided. This listing will be presented by treatment arm and will include: center, subject identifier, age, gender, race, adverse event (SOC, PT, and verbatim term), date of onset, date of resolution, duration, CTCAE grade, seriousness, action taken, outcome, and relatedness. The following AE listings will also be provided using a similar layout:

- Subjects who died during the study
- Subjects with treatment-emergent SAEs
- Subjects with treatment withdrawal due to TEAEs
- Subjects with dose reduction due to TEAEs
- Subjects with dose interruption due to TEAEs

For missing or partially missing dates, imputation will be done according to Section 4.7.1.3. In general, however, adverse events will be assumed to be treatment-emergent unless there is clear evidence (through comparison of partial dates) to suggest that the adverse event started prior to the date of first dose of study treatment or more than 30 days after the date of last dose of study treatment.

4.8.3 Prior and Concomitant Medications

All medications will be collected from the time the subject signs the informed consent form throughout the subject's participation in the study, including a period of 30 days after the subject's last dose of study drug.

For each concomitant medication taken during the trial, the generic component will be identified using the World Health Organization Drug Dictionary (WHO-DD; WHODDB3 ENHANCED JUN 2015).

Medications will be assigned to a time period (prior and/or concomitant) as follows:

- If both the start and stop date exist and are before the first dose date of study drug, the medication will be counted as prior.
- If the start date is on or after the first dose date of study drug, the medication will be counted as concomitant.

- If the start date is before the first dose date of study drug and the stop date is after the first dose date of study drug or the medication is continuing, the medication will be counted as prior and concomitant.
- If the start date is missing and the stop date is before the first dose of study drug, the medication will be counted as prior.
- If the start date is missing and the stop date is after the first dose of study drug or the medication is continuing, the medication will be counted as concomitant.
- If the start and stop dates are missing, the medication will be counted as concomitant.

Summaries showing number of subjects and percentage taking each medication will be provided for each preferred medication term by treatment arm. This will be done separately for prior and for concomitant medications.

A by-subject listing of all prior and concomitant medication data will also be provided.

4.8.4 Clinical Laboratory Evaluation

Local laboratories will be used for laboratory safety evaluations in this study. For a list of the parameters to be evaluated, see [Section 3.2.2](#) of this SAP. Laboratory normal ranges will be provided by the local laboratory. For parameters where an NCI CTCAE v.4.03 scale exists, laboratory results will be graded according to the NCI CTCAE v.4.03 severity grade. For parameters for where an NCI CTCAE v.4.03 scale does not exist, an indicator of whether the value is below, within, or above the normal range will represent severity instead.

All laboratory values will be reported in SI units.

Summaries for each laboratory parameter will be presented by treatment arm and visit. For by-visit summaries, the first non-missing assessment (including repeat assessments) recorded at each visit will be used. Optional laboratory parameters will not be summarized by treatment arm, only listed.

For laboratory values reported as a character value, such as <40, will be transformed into numerical values for summary reasons by following the specified guidelines. If a laboratory value is reported using the less than symbol, '<', 0.10 will be subtracted from the original numeric value. If a laboratory value is reported using the greater than symbol, '>', 0.10 will be added to the original numeric value.

Observed and change from baseline values in laboratory test results will be analyzed using descriptive statistics (except for parameters with categorical results, which will use frequencies and percentages). Change from baseline will only apply to post-baseline assessments for parameters with continuous results. For each lab parameter at each time point, the difference in mean changes from baseline (tivozanib – sorafenib) using a rank-

sum test will be presented along with the 95% CI. Laboratory results will also be summarized by maximum CTC grade or by worst severity, depending on the parameter.

Shifts in grade from baseline to each post-baseline visit will be summarized by treatment arm. The maximum shift (across all visits) from baseline will also be summarized by treatment arm. If needed, significance of shifts in laboratory parameters may be further evaluated within each treatment arm using a generalized McNemar's test.

A by-subject listing of all laboratory data, with abnormal values flagged, will be provided by treatment arm. This listing will include center, patient identifier, age, gender, race, weight, and visit, as well as laboratory reference ranges for each parameter. A similar listing with out-of-reference range values will also be presented.

4.8.5 Vital Signs, Physical Findings and Other Observations Related to Safety

A description of vital sign parameters, ECG results, ECOG performance status, and physical examination can be found in [Section 3.2.2](#) of this SAP.

Summaries for each vital sign parameter will be presented by treatment arm and visit. Observed and change from baseline values in vital sign results will be analyzed using descriptive statistics. Change from baseline will only apply to post-baseline assessments. For each vital sign parameter at each time point, the difference in mean changes from baseline (tivozanib – sorafenib) using a rank-sum test will be presented along with the 95% CI.

Shifts in ECG results, ECOG performance status, and physical examination from baseline to each post-baseline visit (where available) will be summarized by treatment arm. The maximum shift (across all visits) in ECG results and ECOG performance status from baseline will also be summarized by treatment arm.

For systolic blood pressure (SBP) and diastolic blood pressure (DBP), observed values will be classified as markedly abnormal (MA) or not markedly abnormal (NMA). The criteria for determining MA values are as follows:

- SBP ≤ 80 mmHg or ≥ 180 mmHg
- DBP ≤ 50 mmHg or ≥ 105 mmHg

The number and percentage of subjects with MA change values will be summarized for each vital sign parameter by treatment arm and visit.

By-subject listings of vital sign data, ECG results, ECOG performance status, and physical examination results will be provided by treatment arm.

4.8.6 Safety Monitoring Committee

The objectives of the Safety Monitoring Committee (SMC) are to evaluate interim safety data to protect subject welfare and to identify issues and risks, as well as to provide

recommendations regarding study design and conduct. The SMC is composed of six members, all of whom are qualified clinicians in general oncology and renal cell carcinoma and have practical experience conducting and monitoring the safety and efficacy of clinical trials.

A meeting to review cumulative safety data will be planned after the first 50 subjects are randomized to each treatment arm and approximately every 6 months thereafter. There may be additional ad hoc meetings (via teleconference or face-to-face) to evaluate the safety of the subjects, when internal reviews detect safety risks. The frequency may be modified at the discretion of the SMC

After each meeting, the SMC will make recommendations to modify or terminate the trial if there is a safety concern or a significant emerging issue with trial design or conduct. Further details on the Safety Monitoring Committee activities can be found in the SMC Charter.

4.9 Determination of Sample Size

A sample size of 322 subjects (161 subjects per treatment arm) with a total number of 255 events will provide 90% power to detect a statistically significant difference between treatment arms with respect to PFS as assessed by the IRR. The sample size was determined based on the following assumptions:

- The distribution of the PFS for the two treatment arms will be compared by using a log-rank test with one-sided 2.5% significance level.
- The median PFS for subjects receiving sorafenib and tivozanib is 4 months and 6 months, respectively (an increase of 2 months or 50%).
- An equal number of subjects will be assigned to each treatment arm.
- Enrollment will take 15 months.
- The drop out percentage per treatment arm will be 3%.

The first secondary analysis will compare OS between sorafenib and tivozanib by using a stratified log-rank test (stratification factors included in the primary analysis), and two-sided 0.8% significance level (utilizing a Lan-DeMets alpha spending function with an O'Brien Fleming boundary). The interim OS analysis will be done at the time of the final PFS analysis. Assuming a median OS for sorafenib of 12 months and a median OS for tivozanib of 16 months, 245 OS events at the time of final OS analysis and 149 OS events at the time of final PFS analysis will yield approximately 61% power to demonstrate longer OS for tivozanib using the log-rank test, given 2.5% type I error (one-sided). The final OS analysis will be performed when all subjects have been lost to follow-up, have withdrawn consent, have died, or when all subjects in follow-up have been on-study for at least 2 years, whichever occurs first.

The estimate was obtained by using East® statistical package Version 6.3 by Cytel Inc.

4.10 Changes in the Conduct of the Study or Planned Analysis

Protocol Version 4.0 was finalized on 01 October 2018. The administrative change was:

- A previously unspecified data cut-off date for the primary analysis was established to be 04 October 2018 with a targeted number of 242 PFS events.

The PFS event rate continues to decelerate. The study steering committee participated in a conference call on Thursday, August 30, 2018 and unanimously recommended performing the analysis rather than waiting for the additional events to reach the planned 255 PFS events. The reasons given were that the current patients have been on study for at least one year and may not progress for some time, and that the small reduction in events at the time of final analysis was unlikely to materially affect the results.

- The committee was concerned that based on the small number of patients remaining that can contribute to a PFS event, it could take another 5 months or more to get to 255 events.
- Some committee members expressed that based on the safety profiles they have observed during the study, they are keen to provide access to this drug to their patients.

This proposal was presented to the US FDA and they reported the decision is at the discretion of the Steering Committee and the sponsor.

Protocol Version 3.0 was finalized on 13 July 2016. Changes that have statistical implications are:

- the removal of the tertiary objective of exploring the correlation of serum-neuropilin 1 (NPR-1) with clinical activity in subjects randomized to treatment with tivozanib or sorafenib.
- The addition of the tertiary objectives of exploring the relationship between tivozanib and sorafenib drug levels and activity as well as the relationship between the IP drug levels and AEs.

Subjects must begin planned treatment of the study drug within 14 days after randomization.

An exploratory analysis of PFS on the safety population was added to the SAP to account for subjects that were randomized to a treatment arm but not dosed due to external issues.

Best Overall Response criteria was updated to align with the RECIST 1.1 guidelines.

5 REFERENCES

[1] International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use: ICH Harmonised Tripartite

Guideline Statistical Principles for Clinical Trials E9. (1998). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf

[2] New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). (2009). <http://www.eortc.be/Recist/documents/RECISTGuidelines.pdf>

PAREXEL International Electronic Signature Page

This page is the manifestation of the electronic signature(s) used in compliance with PAREXEL International's electronic signature policies and procedures and in compliance with applicable regulations.

UserName: Blair, Anne (blaira)
Title: Senior Biostatistician, GRO
Date: Sunday, 30 September 2018, 08:50 PM GMT Standard Time
Meaning: Document contents authored.

UserName: Roessner, Martin (roessnm)
Title: Corporate VP, Biostatistics, GRO
Date: Monday, 01 October 2018, 12:28 AM GMT Standard Time
Meaning: Document contents reviewed.

UserName: Gerstbrein, Beate (gerstbb)
Title: Principal Medical Writer, GMS
Date: Monday, 01 October 2018, 07:15 AM GMT Standard Time
Meaning: Document contents reviewed.

UserName: Blair, Anne (blaira)
Title: Senior Biostatistician, GRO
Date: Monday, 01 October 2018, 12:50 PM GMT Standard Time
Meaning: Document contents approved.
